DOCUMENT RESUME

ED 427 083 TM 029 466

AUTHOR Helms, LuAnn Sherbeck
TITLE Basic Concepts in Classical Test Theory: Tests Aren't Reliable, the Nature of Alpha, and Reliability Generalization as a Meta-analytic Method.
PUB DATE 1999-01-00
NOTE 16p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; *Meta Analysis; *Reliability; *Scores; *Test Theory; Testing Problems
IDENTIFIERS *Alpha Coefficient

ABSTRACT

This paper discusses the fact that reliability is about scores and not tests and how reliability limits effect sizes. The paper also explores the classical reliability coefficients of stability, equivalence, and internal consistency. Stability is concerned with how stable test scores will be over time, while equivalence addresses the relationship between various forms of a test. Coefficient alpha is one of the common ways to describe the internal consistency of a test. The features that influence coefficient alpha are described, and the new "reliability generalization" meta-analytic technique is also summarized. (Contains 3 figures and 13 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running head: BASIC CONCEPTS IN CLASSICAL TEST THEORY

ED 427 083

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
LuAnn Helms
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Basic Concepts in Classical Test Theory: Tests Aren't Reliable, the Nature of Alpha, and Reliability Generalization as a Meta-analytic Method

LuAnn Sherbeck Helms

Texas A&M University 77843-4225

TM029466

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 21, 1999.

BEST COPY AVAILABLE

Abstract

Too many researchers speak of "the reliability of the test," thus displaying their basic misunderstanding of reliability. This paper explains classical reliability, and the features that influence coefficient alpha, including when it can be negative even though alpha is a variance-accounted-for statistic. The new "reliability generalization" technique is also summarized.

Basic Concepts in Classical Test Theory: Tests Aren't Reliable, the Nature of Alpha, and Reliability Generalization as a Meta-analytic Method

Too many researchers speak of "the reliability of the test," this moronic statement should *never* be used. Reliability is a characteristic of scores or data, and *not* a characteristic of instruments. As Crocker and Algina (1986, p. 144) argued, "...A test is not 'reliable' or 'unreliable.' Rather, reliability is a property of the scores on a test for a particular group of examinees." An IQ test given to a group of adults may yield *reliable scores* on one occasion, but can yield *unreliable scores* when given to a group of children. In classical test reliability theory, reliability is the ratio of true score variability to the observed score variability. Typically, greater score variance leads to greater score reliability. Therefore, a more heterogeneous group of examinees often leads to more variable scores and thus the higher score reliability (Thompson, 1994b).

One should never say a test is reliable or unreliable. "Instead, authors should use language such as 'the scores in our study had a classical theory test-retest reliability coefficient of X.'" as Thompson (1994a) has argued. According to Thompson (1992, p. 436):

> This is not just an issue of sloppy speaking—the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometime lead to a more pernicious outcomes, sloppy thinking and sloppy practice.

This paper explains classical reliability, why reliability is about scores and not about tests, and how reliability limits effect sizes. The classical reliability coefficients of stability, equivalence, and internal consistency will be discussed. Considering that internal consistency is the most often used classical reliability coefficient, the features that influence coefficient alpha will be explored.

Since tests are not reliable, score reliability fluctuates from administration to administration, and must be evaluated in every study. Therefore, the new "reliability generalization" meta-analytic technique is also summarized.

Classical Test Reliability Theory

The True Score Model

Classical test reliability theory is based on the true score model. This model assumes that each person has a true score that equals the actual amount of the characteristic being measured by the test, that would be obtained if there were no errors in the measurement (Kaplan & Saccuzzo, 1989). Measurement error consists of anything that causes a discrepancy between an observed score and a true score (e.g., how the test taker feels that day, the room temperature, the way the directions are given).

If we were to make a distribution of infinite observed scores for repeated testing for the same person, that person's true score would equal the average of the observed scores and the dispersion would represent the distribution of random errors (Kaplan & Saccuzzo, 1989). Since giving an infinite number of testing is impossible, both true score and error cannot be measured directly; they can only be estimated. The standard deviation of this distribution, called the standard error of measurement, tells us about the magnitude of measurement error. By looking at the distribution of observed scores in Figure 1 you can see that distribution A has more measurement error than distribution B.

Insert Figure 1 here

The standard error of measurement is "the standard deviation of a theoretically normal distribution of test score obtained by one person on equivalent tests" (Cohen, Montague, Nathanson, & Swerdlik, 1988, p. 117). Using the following formula, the standard error of measurement can be used to estimate the range within which the individual's true score probably falls. The standard error of measurement is computed by the formula:

$$\sigma_{meas} = s\,(1 - r)^{.5}$$

where σ_{meas} is the standard error of measurement, s is the standard deviation of the test scores, and r is the reliability coefficient (Cohen et al., 1988).

For example if a math test has a standard deviation of 10 and a reliability coefficient of .91, then the standard error of measurement would equal 3. Because 95% of the scores are expected to occur with in + or - $2\sigma_{meas}$, if an individual scored 40 on this test, with 95% confidence we could estimate the subjects true score to be between 34 and 46.

As stated earlier, classical test reliability theory is based on the true score model. This model assumes that each observed score is composed of two parts, the true score and random error. Accordingly, observed score variance is composed of true score variance, the reliable variance, and error variance, the unreliable variance. As portrayed in Figure 2, these two score elements are presumed to be uncorrelated.

Insert Figure 2 here

In classical test theory the reliability coefficient is the ratio of the true score variability to the observed score variability (Kaplan & Saccuzzo, 1989). In other words, the reliability coefficient is the ratio of the reliable variance to the total variance. It is a variance-accounted-for

statistic. In this way measurement and substantive variance partitioning are the same. They are similar because all substantive analyses are in effect regression, that is, they all produce a y-hat and an error score that when added equal the sum of squares total, or observed variance (Dawson, 1997). According to Thompson (1988) *all* classical analytic methods are correlational. In classical test theory the correlation between true scores and observed scores is called the reliability index. The reliability coefficient is the reliability index squared. Many classical reliability coefficients are obtained by using the Pearson r (Pearson product-moment correlation).

$$\text{Pearson } r = [\Sigma (X - \overline{X})(Y - \overline{Y})] / [[[\Sigma (X - \overline{X})^2][\Sigma (Y - \overline{Y})^2]].5]$$

Effect Size

The fact that the reliability coefficient is the ratio of reliable variance to total variance is extremely important, because an effect size is limited by the amount of reliable variance. That is, effect size is limited by the score reliability. The correlation between the scores from two tests can never exceed the square root of the product of the reliability coefficients for the two sets of scores (Thompson, 1994b). If we were to correlate scores from an IQ test with a reliability coefficient of .7 with score from an achievement test with a reliability coefficient of .6, knowing $r^2_{xy} \leq [(\text{reliability of X})(\text{reliability of Y})]^{.5}$, the maximum effect size we could possibly achieve would be .648 for this study.

Reinhardt (1996) reasoned, "if a dependent variable is measured such that scores are perfectly unreliable, the effect size in the study will unavoidably be zero, and the results will not be statistically significant at any sample size.... Prospectively, researchers must select measures that will allow the detection of effects at the level desired; retrospectively, researchers must take reliability into account when interpreting findings" (p. 3). Not only do many researcher neglect to report the reliability coefficients for their data, they may also neglect the fact the reliability

establishes a ceiling for their effect sizes, and conduct studies that could not possibly yield noteworthy effect sizes (Thompson, 1994b). Since effect size is *always* limited by score reliability, all researchers should report and discuss reliability coefficients for their data when interpreting effect sizes. Unfortunately, many researchers neglect to do this (Thompson 1994b; Vacha-Haase 1998).

Types of Reliability Coefficients

Classical test theory provides estimates for at least three types of reliability: stability, equivalence, and internal consistency (Cohen et al., 1988). Each reliability estimate considers *one* source of error: either error in test occasions, test forms, or in items.

The Coefficient of Stability (also called test-retest reliability coefficient) is concerned with how stable observed test scores will be over time. This coefficient is obtained by correlating pairs of scores obtained from the same subjects on two different administrations of the same test. A high reliability coefficient suggests that measurements given on two occasions will yield relatively the same scores.

Coefficient of Equivalence (also called alternate-form or parallel-form reliability coefficient) addresses the degree of the relationship between various forms of a test. Half the subjects are given test A and half the subjects are given test B, and when finished the subjects then take the other test. By correlating the scores obtained on test A to the scores obtained on test B an equivalence reliability coefficient is obtained. A high reliability coefficient suggests that the parallel forms could be used interchangeably with confidence.

It is often difficult to develop two tests or to test and retest the same subjects. Therefore may researchers prefer to calculate reliability based on the scores obtained from a single administration of the items of the test. This reliability coefficient is called the Coefficient of

Internal Consistency. A high reliability coefficient suggests that the items are homogeneous (the same kind) with respect to statistical characteristics of interest. There are several ways to compute the internal consistency coefficient. The most commonly used methods include: Split-half reliability estimates, Kuder-Richardson 20, and Cronbach's Coefficient Alpha.

Split-half coefficients are the correlations between scores on two halves of the same test. According to Cohen et al. (1988), this coefficient is calculated by first dividing the test in to equivalent halves. There are several ways to divide the test. However, simply dividing the test in half (top versus bottom) is not recommended. Often the test is divided by putting the odd items in one group and the even items in another, or the items are randomly assigned to two groups. Second, compute a Pearson r between score on each half. Third, adjust the half-test reliability using the Spearman-Brown formula. Generally, but not always, reliability increases as test length increases, providing that the additional items are equivalent with respect to the content and the range of difficulty of the original items (Cohen et al., 1988). The coefficient obtained in step 2 needs to be adjusted because it only represents half of the test. The formula for this adjustment is as follows:

$$r_{SB} = 2r / (1 + r),$$

where r_{SB} is the reliability adjusted by the Spearman-Brown formula and r is the Pearson r. Split-half coefficients are not recommended when the two halves of the test have unequal variances (Kaplan & Saccuzzo, 1989).

Coefficient Alpha

Since there are many ways to split a test and different splits may yield contradictory results, coefficient alpha is often preferred to split-half reliability estimates. Crocker and Algina (1986, p. 142) defined alpha as, "the mean of all possible split-half coefficients that are

calculated." Coefficient alpha provides the lowest estimate of reliability that can be expected (Kaplan & Saccuzzo, 1989). According to Reinhardt (1996), alpha can be interpreted as "the lower bound estimate of the proportion of variance in the test scores explained by common factors underlying item performance." Coefficient alpha is usually determined by using one of two very similar formulas: the Kuder-Richardson formula #20 (Kuder & Richardson, 1937) or Cronbach's Alpha (Cronbach, 1951). The formula for the KR-20 is:

$$KR20 = k/(k-1) * [\ 1 - (\Sigma pq^2 / \sigma_T^2\ )],$$

The formula for the Cronbach's Alpha is:

$$alpha = k/(k-1) * [\ 1 - (\Sigma \sigma_k^2 / \sigma_T^2\ )]$$

k = number of items
p = percent of persons answering the item correctly
q = percent of persons answering the items incorrectly
σ_k^2 = variance of one item
σ_T^2 = variance of the total test scores.

Notice that the formulas are identical with one exception, how they compute the sum of the item variances. As you can see in Figure 3 the KR-20 formula makes calculating item variances simpler. But, the KR-20 can only be used only with dichotomously scored data (i.e., data that are scored right or wrong).

Insert Figure 3 here

The formulas show that coefficient alpha is affected by item difficulty, sum of item variances, and total test score variance. Reinhardt (1996) provided an excellent review of how each of these factors affect coefficient alpha and demonstrated that total test score variance has the biggest effect on coefficient alpha. The smaller the total test score variance, the smaller

coefficient alpha will tend to be. When there is no variability in total test score, then $\sigma_T^2 = 0$ and it is impossible to compute coefficient alpha. If the variability in total test scores is less than the sum of the item variances, then coefficient alpha will be negative. According to Reinhardt, total test score variance is maximized when "half of the examinees earn the lowest possible total score and half earn the highest possible total score" (p. 9). This arrangement creates maximum deviation from the mean test scores and will produce maximum total test score variance.

The total test score variance is greatly effected by how homogeneous or heterogeneous are the group of examinees. For example, if a test is given to a group of graduate students with the same major, background, and grade point average, they are more likely to answer the questions the same way, thus decreasing variability in total test score and decreasing coefficient alpha. However, if the same test is given to a heterogeneous group of high school students, the variability in their answer will most likely be greater, and hereby increasing coefficient alpha. This is another demonstration that tests are not reliable, because the same test given to different groups can yield dramatically different reliability coefficients.

Reliability Generalization Technique

Since tests are not reliable and score reliability fluctuates from administration to administration, Vacha-Haase (1998) proposed a new method called "reliability generalization," as a way to explore score reliability across studies. Take care not to confuse reliability generalization with generalizability theory. Generalizability theory is a modern reliability estimation procedure that is often compared to classical reliability theory (Eason, 1991; Thompson, 1992; Thompson & Crowley, 1994). According to Vacha-Haase, reliability generalization is meta-analytic technique that characterizes (a) the typical score reliability for a given measure across administrations, (b)

the variability of score reliability for a given measure, and (c) the sources of variability in score reliability across studies. Vacha-Haase (1998) stated:

> Reliability generalization is a potentially powerful method with which to characterize and explore variance in score reliability. The potentials of the method are honored in the editorial policies of this journal [Educational and Psychological Measurement], which now encourage the submission of manuscripts employing reliability generalization.

Vacha-Haase (1998) applied the reliability generalization method to the Masculine and Feminine Scales of the Bem Sex Role Inventory (BSRI). Her reliability generalization procedure included four steps. The first step was gathering articles. She found 628 articles that used the BSRI. Second, the articles where reliability coefficients were reported in a meaningful manner were identified. Only 57 out of the 628 BSRI articles met this criteria. Third, articles were coded based on types of reliability coefficients, form length and format, language test was administered in, and gender and student status of participants. Finally, the data were analyzed across studies to characterize (a) mean score reliability, (b) the variability in score reliability, and (c) the study features that tend to predict variability in score reliability. Vacha-Haase's results indicated that the reported reliability coefficients were fairly variable across the studies using the BSRI, thus providing a demonstration that reliability does not inure to tests but rather to scores.

References

Cohen, R. J., Montague, P., Nathanson, L. S., & Swerdlik, M. E. (1988). Psychological testing: An introduction to tests and measurement. Mountain View, CA: Mayfield Publishing.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winson.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of test. Psychometrika, 16, 297-334.

Eason, S. (1991). Why generalizability theory yields better results that classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press

Kaplan, R. M., & Saccuzzo, D. P. (1989). Psychological testing: Principles, applications, and issues. Pacific Grove, CA: Brooks and Cole.

Kuder, G. F., & Richardson, M.W. (1937). The theory of estimation of test reliability. Psychometrika, 2, 151-160.l

Thompson, B. (1988, November). Common methodology mistakes in dissertations: Improving dissertation quality. Paper presented at the annual meeting of the Mid-South Educational Research Association, Louisville, KY. (ERIC Document Reproduction Service No. ED 301 595)

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1994a). Guidelines for authors. Educational and Psychological Measurements, 54, 837-847.

Thompson, B. (1994b, April). Common methodology mistakes in dissertations, Revisited. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 368 771)

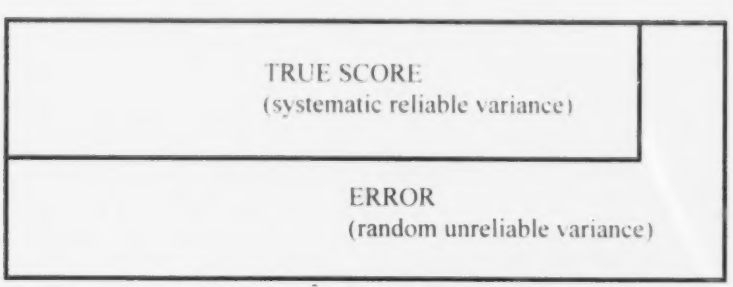
Thompson, B. & Crowley, S. (1994, April). When classical measurement theory is insufficient and generalizability theory is essential. Paper presented at the annual meeting of the Western Psychological Association, Kailua-Kona, Hawaii. (ERIC Document Reproduction Service No. ED 377 218)

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

Figure 1. Two distributions of observed scores.

Figure 2. Venn diagram displaying true score model.



OBSERVED SCORE VARIANCE

Figure 3. Computation of item variance using Kuder-Richardson 20 and Cronbach's alpha

Kuder-Richardson 20 item variance

People	Item Scores
1	0
2	0
3	0
4	1
5	1
6	1
7	1
8	1

Var pq = (0.625) (0.375) = 0.234375

Cronbach's alpha item variance

Scores	- Mean	= Deviation	Dev. squared
0	0.625	-0.625	0.390625
0	0.625	-0.625	0.390625
0	0.625	-0.625	0.390625
1	0.625	0.375	0.140625
1	0.625	0.375	0.140625
1	0.625	0.375	0.140625
1	0.625	0.375	0.140625
1	0.625	0.375	0.140625
			1.875 = SOS

Var = SOS / n

Var = 1.875 / 8 = 0.234375